

General human activity patterns

Anders Mollgaard,¹ Sune Lehmann,² Joachim Mathiesen¹

¹Niels Bohr Institute, University of Copenhagen,
Blegdamsvej 17, 2100 Copenhagen, Denmark.

²Technical University of Denmark,
Anker Engelundsvej 1, 2800 Kgs. Lyngby, Denmark.

We investigate the dynamics and interplay between human communication, movement, and social proximity by analyzing data collected from smartphones distributed among 638 individuals. The main question we consider is: to what extent do individuals act according to patterns shared across an entire population? Based on statistics of the entire population, we successfully predict 71% of the activity and 85% of the inactivity involved in communication, movement, and social proximity. We find that individual level statistics only result in marginally better predictions, indicating a high degree of shared activity patterns across the population. Finally, we predict short-term activity patterns using a generalized linear model, which suggests that a simple linear description might be sufficient to explain a wide range of actions, whether they be of social or of physical character.

Introduction

Human behavior is inherently difficult to predict and not least to model [1]. Recent advances in data collection techniques [2, 3, 4, 5, 6], however, have made it increasingly feasible to build and test models of human behavior against empirical data. In particular, there has been progress on understanding social networks, both in terms of predicting link formation [7, 8] and the dynamics of link activity [9, 10, 11]. Similarly there has been progress on developing models for collective human dynamics, especially in the field of online attention [12, 13, 14]. Models of individual human dynamics have been characterized mainly by two approaches. Either they seek to reproduce statistical properties of general human activity [15, 16, 17, 18, 19], but avoid the question of predictability. Else they focus on the predictability of human mobility, but exploit individual patterns that do not apply in general [20, 21, 22, 23]. Still unexplored is the feasibility of predicting individual dynamics based on general patterns. Also, the concept of activity is broader than the concept of mobility, since it may include communication, face-to-face interaction, working, sleeping, etc. By studying the interactions among different human activities, we obtain a richer description of human dynamics. A characterization of general activity patterns can provide a first step towards a comprehensive bottom-up description of collective human phenomena.

In this paper we analyze the dynamics of calling, texting, movement, and social proximity. The data is collected from smartphones distributed among 638 students at a large European University [6] (see Materials and Methods for prior filtering). Custom software was installed on the smartphones, which allowed us to gather de-identified data regarding calls, texts (SMS), GPS and Bluetooth from consenting students. From the GPS information we derive movement, and from the Bluetooth we derive social proximity among the test subjects (see Supplementary Material Section S1). By binning the data collection period of 18 months into bins of 15

minutes duration, we are able to produce time series, $x_i^{(u)}(t) \in \{0, 1\}$, describing the activity of user, u , in channel, i , at time, t . Note that our representation is binary, such that each channel is represented by either activity or inactivity. In Fig. 1 we show an example of the activity dynamics for a single user during a single week (spikes) along with the average weekly activity across the full data set. The weekly average reveals an underlying circadian pattern, which therefore mediates a correlation between the individual activity spikes. However, we also expect the activity spikes to interact, thereby bringing further structure to the dynamics. Below we investigate the questions: Are there further activity patterns on top of the circadian ones? Are these patterns shared among all individuals? If so, what characterizes them?

Results

Predicting activity

If human activity follows distinct patterns, then the future activity of an individual may be predicted with some certainty. Our first task is to quantify the degree of predictability inherent to our data set. We define a predictive pattern to be a set of consecutive time bins of total length Δt_h , which is separated from a future activity state by a time t_f . In Fig. 2A we show an example of such a predictive pattern for the case of $\Delta t_h = 45$ min and $t_f = 30$ min. Person Z is currently on the move talking to someone on the phone and we know that another call was made 15 minutes ago, but apart from that there has been no other activity during the last 45 minutes. The task is to make the best possible prediction of the future state based on the observed history.

Before moving on to the actual predictions, it is important to note that the data set has a strong class imbalance. In particular, calls are present in only 2.7% of the time bins, texts in 7.3%, movement in 9.4%, and social proximity in 8.8%. The fraction of correct predictions is therefore not a useful measure of predictability, since guessing "no activity" in all bins will

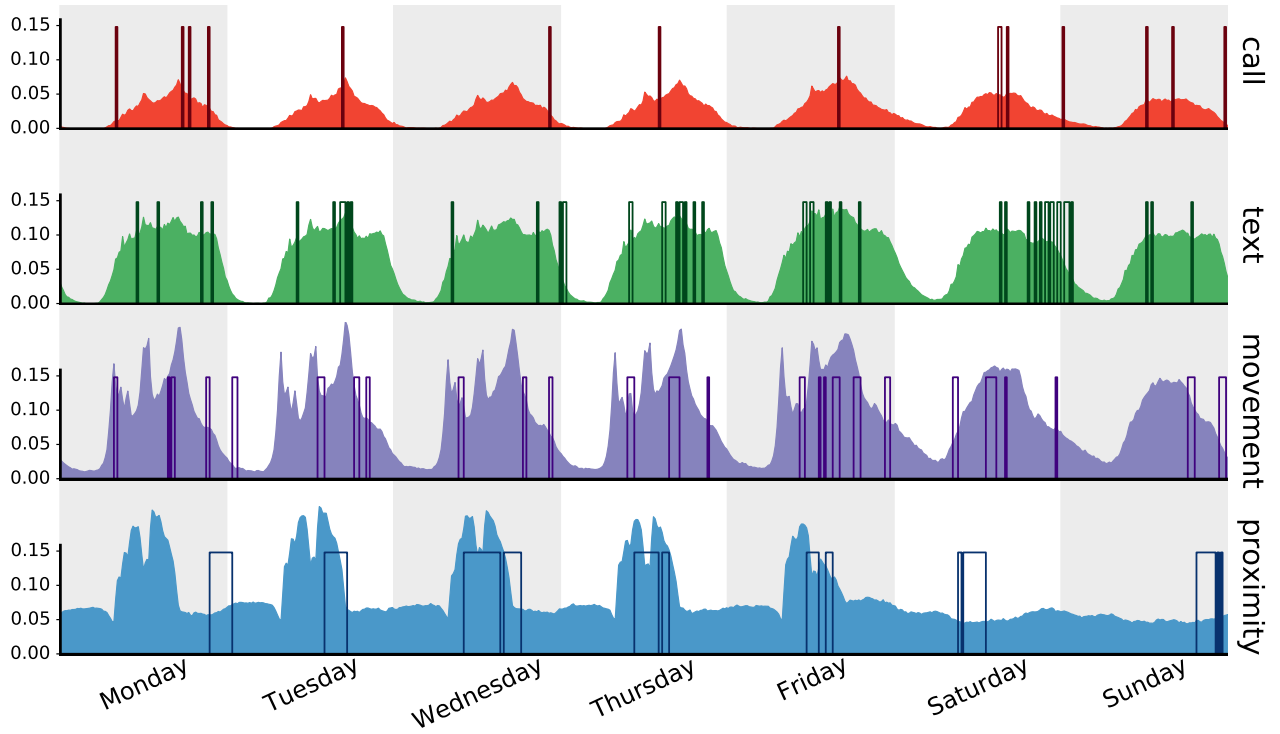


Figure 1: **Activity example.** We show the activity of a single user during one week (spikes), along with the average weekly activity across all users and all weeks. From the top to the bottom panel we show the activity of calls (red), texts (green), movement (purple), and social proximity (blue). The underlying circadian pattern is obvious, but there is also additional structure to be found among the activity spikes.

yield a very high score. We address this bias by instead measuring the *informedness* [24] of our predictions, $I = R_{11} + R_{00} - 1$. Here R_{11} and R_{00} are the fractions of predictive patterns with respectively active and inactive future states that have been correctly predicted. Consistently predicting either state or guessing at random will result in an informedness of 0. For instance, by guessing "no activity" for all the predictive patterns we will succeed in predicting all future inactivity ($R_{00} = 1$) and fail in predicting any future activity ($R_{11} = 0$), thereby obtaining an overall informedness of 0. Note that the measure of informedness applies to the individual activity channel, so for each Δt_h and t_f there will be four different numbers of informedness.

So how do we optimize this informedness? Let N be the size of the data set and let n_i be the number of predictive patterns with future activity in channel i . Suppose that we are making a prediction for a specific predictive pattern, which has a probability p_i of future activity in channel i according to some model. Then, the expected informedness is maximized according to the following decision rule (see Supplementary Material, Section S2):

$$x_{i,\text{predict}} = \begin{cases} 1, & p_i > \frac{n_i}{N}. \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The ratio n_i/N is the probability of future activity in channel i for a random predictive pattern. The rule therefore tells us to predict the future state (active/inactive) that is favored by the model as compared to random.

In order to apply the above decision rule we need a model to provide us with p_i . Let us return to the predictive pattern example shown in Fig. 2A and focus on the task of estimating the probability of activity in the movement channel. For now let us assume that we train our expectations on the predictive patterns of everyone grouped together. This data set includes $N \sim 13.5$ million predictive patterns and $n_{\text{movement}} \sim 1.2$ million of these have movement activity at $t_f = 30$ min, i.e. a ratio of 9%. However, if we focus solely on the predictive pattern presented in Fig. 2A, then the ratio is instead 45%. These 45% represents our best estimate for

the probability of movement in the future state given the particular pattern. According to the decision rule in Eq. (1) we optimize the informedness by anticipating movement at $t_f = 30$ min, since $p_{\text{movement}} = 0.45$ is greater than $n_{\text{movement}}/N = 0.09$. Note that the full information of the predictive pattern is exploited in making this prediction, so it is not possible to get a better prediction with any other model trained on the same data. The only limitation is the size of the data set, which must be large enough to estimate p_i with sufficient precision.

Let us start by showing that our data set is sufficiently large for the case of $\Delta t_h = 45$ min and $t_f = 15$ min. For that purpose, we reduce artificially the size of our data by using a subset only. Within the subset, we repeatedly apply the decision rule Eq. (1) to obtain predictions for all the future states and these can then be compared with the actual values to compute the informedness. In Fig. 2B we show the dependence of the informedness on the size of the data set. We find that the informedness converges in all four channels before the full data is used. However, for longer predictive patterns, we do not observe full convergence (see Supplementary Material, Section S3) indicating that $\Delta t_h = 45$ min is the longest predictive pattern for which we can compute the upper bound on the informedness. In Fig. 2C, we show the dependence of the informedness on the length of the predictive pattern. The connected data points correspond to true upper bounds on the informedness, while the unconnected data points are limited by statistics. Here we use the full size of the data set and fix $t_f = 15$ min. The graph clearly illustrates that the system has a memory, since the informedness improves as more of the past is included (when not limited by statistics). Predictive patterns of length $\Delta t_h = 45$ min allow us to successfully predict 71% of the activity and 85% of the inactivity in the four channels for the near future, $t_f = 15$ min.

In the analysis above, the predictive patterns from everyone are included in the statistics such that patterns of individuals are not visible. Therefore, the common data set might be seen as a general representation of a human being without distinct characteristics. We saw that the

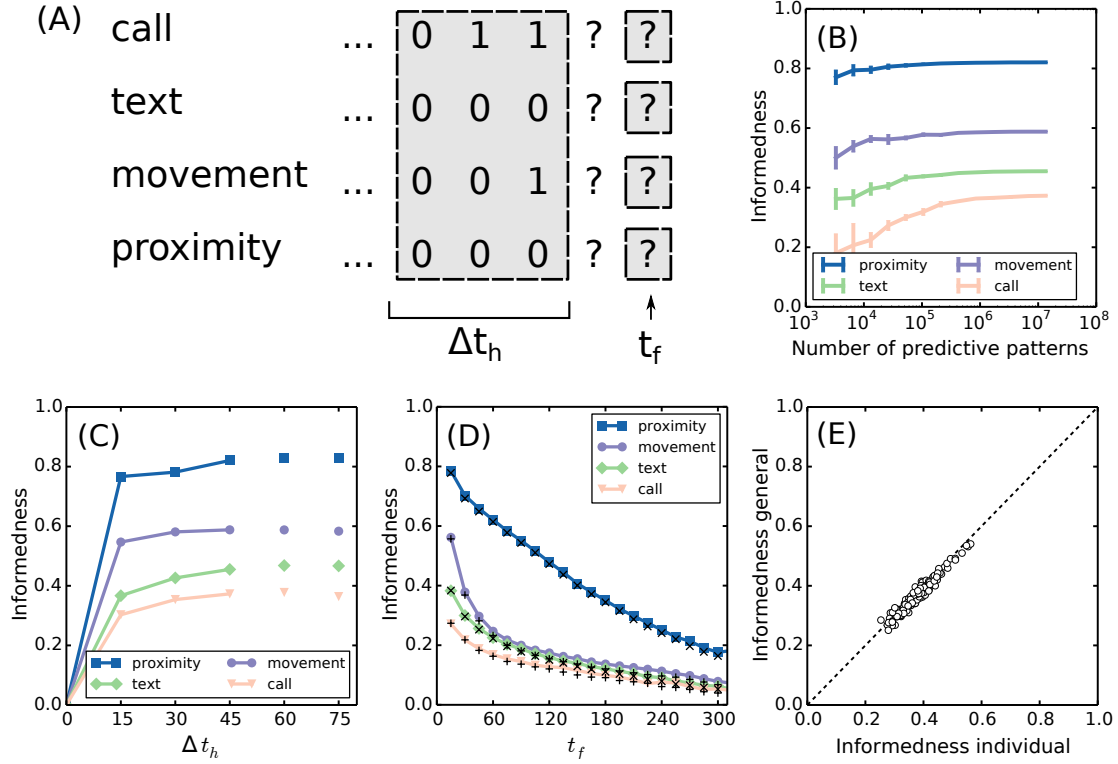


Figure 2: Predicting future activity. (A) We visualize the task of predicting future activity based on a four dimensional time series of length Δt_h . Each time bin has a width of 15 minutes and 0/1 represents activity/inactivity. The probability of activity at $t = t_f$ may be determined from the statistics of similar predictive patterns in the data set. For example, the particular predictive pattern shown here has a 45% probability of activity in the movement channel at $t_f = 30$ min. (B) The data set needs a certain size in order to make accurate probability estimates. Here we show the informedness of the predictions at $t_f = 15$ min based on a predictive pattern of length $\Delta t_h = 45$ min and varying data set sizes. Full convergence is obtained around 10^7 predictive patterns, which means that the data set is sufficiently large for this task. (C) Using the full data set, we then make predictions for $t_f = 15$ min and vary the length of the predictive pattern. By increasing the memory, we also increase the informedness of our predictions, which is clear from the connected markers. The disconnected markers at $\Delta t_h = 60$ min and $\Delta t_h = 75$ min are limited by statistics, meaning that the true upper bound on informedness is not obtained. (D) We then fix the length of the predictive patterns to $\Delta t_h = 45$ min and vary t_f . The lines represent the population average of the informedness for predictions that are based on individual activity patterns. The 'x' and '+' markers represent the population average of the informedness for predictions that are based on common activity patterns. The common patterns almost explain the full information of the individual patterns, which tells us that the activity patterns are general within our population. (E) We expand the population average in terms of individual data sets for the case of movement and $t_f = 30$ min (second predictive pattern from left, purple line, previous figure). The horizontal axis shows the informedness of predictions based on individual predictive patterns, while the vertical axis shows the informedness of predictions based on general predictive patterns. Both measures vary across individuals, but they do so in almost perfect agreement.

near future ($t_f = 15$ min) of the general human could be predicted with a high precision. Now we will have a look at individual data sets and compare the predictions of personal patterns to the predictions of the general patterns. The data on individuals are only a small fraction of the full data set. In the following, we therefore restrict the analysis to persons with at least $3 \cdot 10^4$ predictive patterns, which leaves us with 139 individuals. For each individual data set, we perform the same analysis as above for varying t_f , but with the length of the predictive pattern fixed to $\Delta t_h = 45$ min. We then average over the informedness obtained for the individual data sets, see Fig. 2D. As expected, we find that the predictions are best for the near future. There is a sharp drop in the informedness over the first 60 minutes followed by a linear decay.

Next we investigate the informedness of predictions based on common activity patterns. For each individual, we first sample a subset of predictive patterns from the common data set, which has the same statistics as the individual data set. In particular, we sample such that any predictive pattern occurs exactly one time less in the common data sample than in the individual sample, because this ensures similar statistics for the prediction step. Note that even though the statistics of the predictive patterns are identical, the statistics of the future states are not. By applying the decision rule Eq. (1) to the sample, we might therefore get another set of predictions. The predictions of the common data are tested and the informedness is computed. The average informedness across all individuals is shown in Fig. 2D with 'x' and '+' markers. Somewhat surprisingly, the predictions of the general patterns almost fully match the predictions based on individual patterns. This shows that individual characteristics do not change our predictions - or to put it another way - activity patterns are general across the population. Note that the occurrence of the different patterns is not the same across individuals. Some individuals, for instance, make a lot more calls than others and therefore have more patterns with many calls. But when we see the same predictive pattern for two individuals, then it is the same future prediction that optimizes our informedness. In Fig. 2E we plot, for

each individual, the informedness of the individual patterns against the informedness of the general patterns for future mobility at $t_f = 30$ min. This corresponds to the second purple disc from the left in Fig. 2D, but now without performing the population average. We see that the predictability varies a lot among the individual data sets, but the performance of the individual- and general predictive patterns scale together. In many cases the general predictions even perform better than the individual predictions. This is an artifact of small data set sizes, which means that those predictions are trained on too little data to make optimal predictions. It is therefore by chance that the general patterns sometimes outperform the individual patterns, and indeed for $\Delta t_h = 15$ min, where the statistics are much better, we find that the individual patterns outperforms the general patterns for almost all data sets. Finally, we note that the informedness at $t_f = 15$ min is strongly correlated to the informedness at $t_f > 15$ min. It drops from $C \sim 1$ to $C \sim 0.5$ on the time scale of one hour (calls) to several hours (texts, movement, proximity). Therefore, people who are more predictable on short time scales are also more predictable on long time scales.

Parametric model

In the previous section we exploited the size of the data set to perform separate statistics on all the possible predictive patterns. This nonparametric modeling approach is the optimal choice in terms of predictability, but it does not help us understand the underlying patterns. Furthermore, it is not possible to apply this method for very long predictive patterns and many activity channels, since the number of possible predictive patterns grows exponentially with both. To incorporate a better time resolution, a longer history, and more activity channels, one needs to apply a parametric approach. A parametric approach will necessarily make assumptions about the correlations in the data, which will never be perfectly accurate, but such a model also allows us to interpolate knowledge between different parts of input space. This allows us to

train on much less data. More importantly, we may learn about the central dynamics from the performance of simple models. Let us first consider a very simple model without any parameters; namely one which assumes that the future state is equal to the current one. We label this model ‘Inertia’ and present the informedness of its predictions in Fig. 3 (discs). By comparing this result to that of the nonparametric model (squares), we conclude that the predictability of the near future for movement and proximity is mainly based on inertia. For predictions further into the future and for calls/texts, the dynamics is more complicated. We therefore introduce another model, which keeps things simple, but allows training to the history, namely the ‘logit model’. The logit model is a generalized linear model, which feeds a weighted sum of the input to a sigmoid in order to get a probability estimate for the state of a binary system

$$p_i = \sigma(\mathbf{w}_i \cdot \mathbf{x}_{\text{history}}),$$

$$= \frac{1}{1 + \exp(-\mathbf{w}_i \cdot \mathbf{x}_{\text{history}})}.$$

Here $\mathbf{x}_{\text{history}}$ is a vector representing the activities and inactivities in a predictive pattern and \mathbf{w}_i is a weight vector, which is optimized within a training set according to L1 regularization [25]. We have used 75% of the data set for training (chosen at random) and tested the model on the remaining 25%. Note that the decision rule in (1) again is applied to maximize the informedness, but now with p_i determined from the linear model. The informedness of the predictions are shown in Fig. 3 (triangles). Somewhat surprisingly, the linear model almost matches the upper bound on informedness represented by the nonparametric model. A further analysis shows that the difference in informedness between the two models mainly arise from histories that are characterized by current inactivity, but with activity in the past. Here the non linear model tends to underestimate the probability of future activity; possibly because it would otherwise overestimate the probability of future activity for histories with activity both in the past and present.

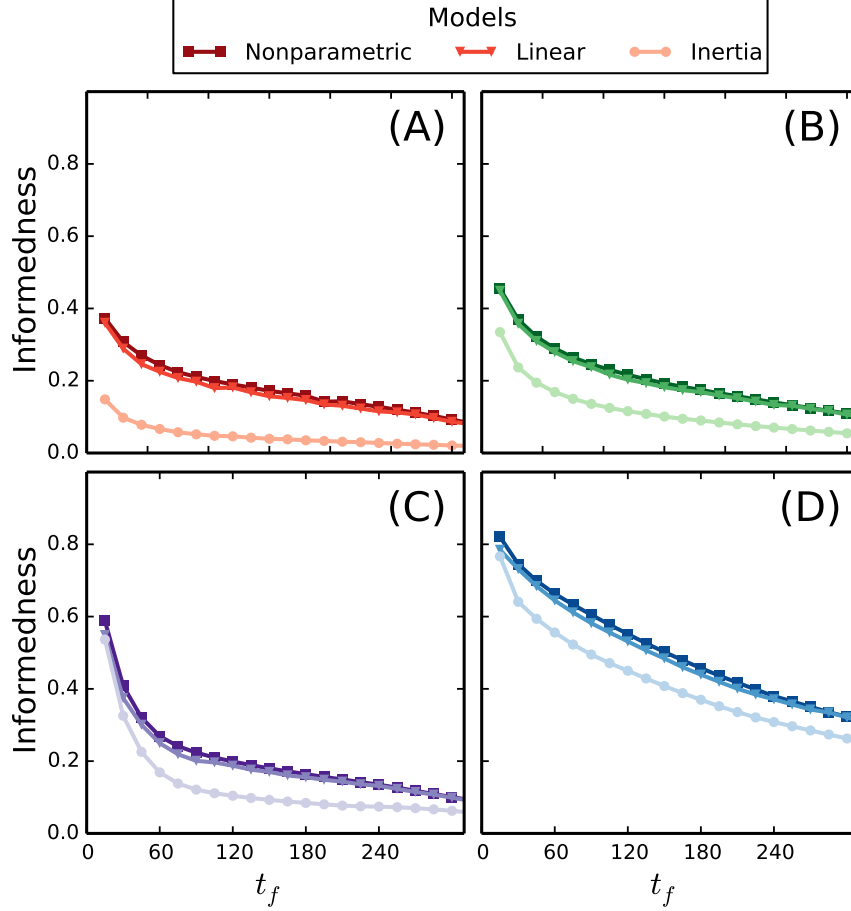


Figure 3: **Model comparison.** We show the informedness of our predictions regarding future activity in the respective channels: call (A), text (B), movement (C), and proximity (D). The horizontal axis shows the reach of the predictions into the future in minutes. The informedness of predictions from three different models are presented. "Nonparametric" (squares) is the label for the nonparametric model that was also presented in Fig. 2, here using a history length of $\Delta t_h = 45$ min. "Inertia" (discs) is a simple model which assumes that the future continues in the same state as the current one. We see that the near future predictions of movement and proximity is dominated by this information. "Linear" (triangles) labels the predictions of a generalized linear model. Note that the informedness of the linear model almost matches the upper bound represented by the nonparametric model.

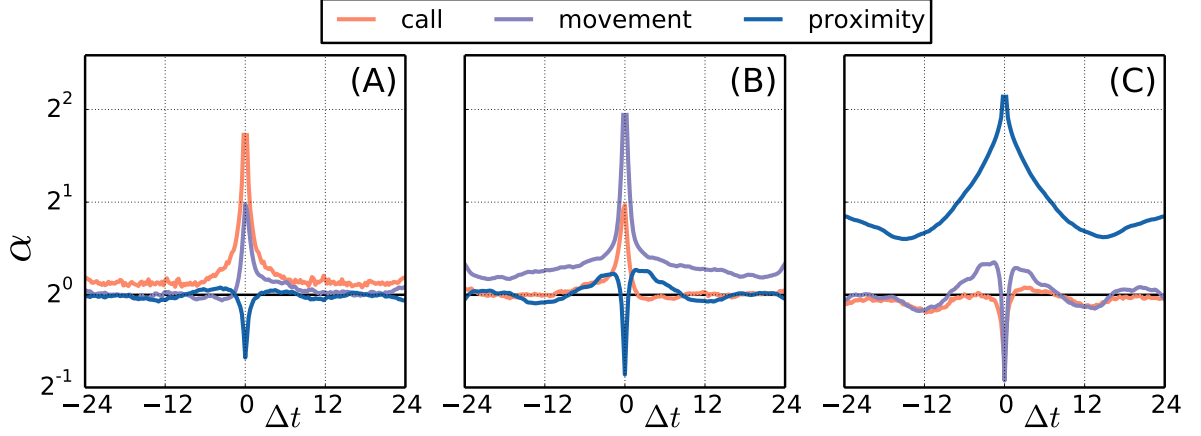


Figure 4: **Activity correlations.** In (A)-(C) we show the increase in activity triggered by respectively calling, movement, and social proximity. The horizontal axis spans 24 hours backward and forward in time and the vertical axis gives the factor of increased activity on a logarithmic scale. For example, the purple line in (A) tells us that movement is enhanced by a factor 2 at the time of a call and significantly increased for several hours after. More generally, all channels are subject to self-enhancement effects, and significant cross-correlations are observed for all combinations of activities.

Correlations

The success of the simple linear model tells us that the general dynamics foremost is controlled by linear correlations. Note that the correlations are partly due to circadian patterns, since we tend to be active in all four channels during daytime and inactive in all four channels during nighttime. We stress that here we investigate correlations mediated by interactions, rather than by circadian cycles. First we compute the probability of activity in channel j a time Δt after observing activity in channel i , $P_{ij}(\Delta t)$. Since the behavior of this function is influenced by circadian cycles, we compute a reference probability function, $Q_{ij}(\Delta t)$ that contains the same circadian pattern, but which is unconditional on activity in channel i (see Supplementary Material, section S4). The ratio of the two, $\alpha_{ij}(\Delta t) = P_{ij}(\Delta t)/Q_{ij}(\Delta t)$, measures the increase of activity in channel j conditioned on activity in channel i excluding the effects of time and weekday. In Fig. 4A-4C we plot $\alpha_{ij}(\Delta t)$ for $i, j \in \{\text{call, movement, proximity}\}$. In Fig. 4A we

see the information gained by observing a call at time $\Delta t = 0$. We notice that the probability of another call 15 minutes after the first is increased by a factor of 3.3; a self-enhancement that is visible across several hours. Similarly, we find that the probability of movement is increased by a factor of 2.0 in the vicinity of a call and with an asymmetric drop off that indicates an arrow of causality mostly pointing from call to movement. Finally, we find that the probability of social proximity is reduced by a factor 0.6 in the vicinity of a call and is slightly increased prior to a call. In 4B we find that movement also shows self-enhancement, but with a peak value of 3.8 and a slightly slower decay. It is likewise found that the probability of social proximity is reduced by a factor 0.6 while moving, but that it is enhanced a few hours before and after moving. In 4C we see that the self-enhancement of social proximity peaks at a value of 4.4 and decays on a time scale that surpasses 24 hours. This time scale is probably not representative of general social proximity, but is more likely to be an artifact of our measurements being restricted to a specific group of people that interact strongly in five day intervals (weekdays) and less so during weekends and holidays. The qualitative effects of texting (not shown) is very similar to calling. The general self-enhancement observed for all four channels is in agreement with previous research stating that humans are characterized by long periods of inactivity followed by bursts of activity [16, 15].

Concluding remarks

We have introduced a framework for analyzing human activity patterns along with a decision rule that allows us to make optimal predictions in terms of informedness. We find that individuals to a large extent act according to a general set of activity patterns, thereby allowing us to predict individual activity with a high precision from general patterns. We believe that our conclusions based on four selected variables carries over to other variables related to human activity. Furthermore, we found that the optimal informedness of the full patterns were almost

matched by a much simpler generalized linear model. We conclude that the self-enhancements and cross correlations presented in Fig. 4 provide an almost complete description of the general dynamics in the system to the extent of the available information.

We should state that our analysis is subject to a number of limitations. First of all, we have considered only four variables. It would be interesting to include other variables such as sleeping, talking, online activity, etc. Secondly, our conclusions are based on a selected population and should be checked on other data sets before one can claim true universality in activity patterns. Finally, the true upper bounds on informedness has only been computed for resolutions of 15 minutes and history information below one hour. This restriction is necessary, since the number of predictive patterns needed to train the nonparametric model grows exponentially with time resolution, length of history, and number of variables. We therefore need to explore other approaches such as the generalized linear model, which almost match the nonparametric model in performance, while also allowing the extension to better time resolutions, longer histories, and more activity channels.

Materials and Methods

The original data set involved 752 participants. We arrive at the 638 participants by filtering out individuals with less than 100 bins of activity in either of the four channels. The average number of active bins in the filtered data is 577 for calls, 1572 for texts, 2021 for movement, and 1894 for social proximity. The filtering is imposed to assure that the smartphone has been used as a main device. The data collection app and the data collection process is described in detail in [6]. The step from raw data to activity signals is described in Supplementary Material, section 1. The activity signals are made public as a JSON file on the website.

References

- [1] D. J. Watts, *Everything is obvious: Once you know the answer* (Crown Business, 2011).
- [2] M. C. Gonzalez, C. A. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
- [3] M. Salathé, *et al.*, A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences* **107**, 22020–22025 (2010).
- [4] J. Stehlé, *et al.*, High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one* **6**, e23176 (2011).
- [5] N. Eagle, A. S. Pentland, D. Lazer, Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences* **106**, 15274–15278 (2009).
- [6] A. Stopczynski, *et al.*, Measuring large-scale social networks with high resolution. *PloS one* **9**, e95978 (2014).
- [7] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks. *Journal of the American society for information science and technology* **58**, 1019–1031 (2007).
- [8] L. L., T. Zhou, Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* **390**, 1150 - 1170 (2011).
- [9] J. Ugander, L. Backstrom, C. Marlow, J. Kleinberg, Structural diversity in social contagion. *Proceedings of the National Academy of Sciences* **109**, 5962–5966 (2012).
- [10] J. Leskovec, L. Backstrom, J. Kleinberg, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2009), pp. 497–506.

- [11] L. Weng, F. Menczer, Y.-Y. Ahn, Virality prediction and community structure in social networks. *Scientific reports* **3** (2013).
- [12] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* **105**, 15649–15653 (2008).
- [13] J. Mathiesen, L. Angheluta, P. T. Ahlgren, M. H. Jensen, Excitable human dynamics driven by extrinsic events in massive communities. *Proceedings of the National Academy of Sciences* **110**, 17259–17262 (2013).
- [14] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, C. Faloutsos, *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2012), pp. 6–14.
- [15] A. Vázquez, *et al.*, Modeling bursts and heavy tails in human dynamics. *Physical Review E* **73**, 036127 (2006).
- [16] Y. Wu, C. Zhou, J. Xiao, J. Kurths, H. J. Schellnhuber, Evidence for a bimodal distribution in human communication. *Proceedings of the national academy of sciences* **107**, 18803–18808 (2010).
- [17] R. D. Malmgren, D. B. Stouffer, A. E. Motter, L. A. Amaral, A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences* **105**, 18153–18158 (2008).
- [18] X.-P. Han, T. Zhou, B.-H. Wang, Modeling human dynamics with adaptive interest. *New Journal of Physics* **10**, 073010 (2008).

- [19] M. Karsai, K. Kaski, A.-L. Barabási, J. Kertész, Universal features of correlated bursty behaviour. *Scientific reports* **2** (2012).
- [20] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).
- [21] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, L. Bengtsson, Approaching the limit of predictability in human mobility. *Scientific reports* **3** (2013).
- [22] E. L. Ikanovic, A. Mollgaard, From a to b: A new approach to the limits of predictability of human mobility patterns. *arXiv preprint arXiv:1608.06419* (2016).
- [23] A. Cuttone, S. Lehmann, M. C. González, Understanding predictability and exploration in human mobility. *arXiv preprint arXiv:1608.01939* (2016).
- [24] D. M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation (2011).
- [25] F. Pedregosa, *et al.*, Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).